Journal of Geographic Information and Decision Analysis, vol. 2, no. 2, pp. 18-30, 1998

Spatial Interpolation and its Uncertainty Using Automated Anisotropic Inverse Distance Weighting (IDW) - Cross-Validation/Jackknife Approach

Maciej Tomczak
Great Lakes Institute for Environmental Research, University of Windsor, Windsor, Ontario N9B 3P4, Canada
Golder Associates Ltd., 2825 Lauzon Parkway, Suite 14, Windsor, Ontario N8T 3H5, Canada
tomczak@sprint.ca
tomczak@rocketmail.com

ABSTRACT **In order to estimate rainfall magnitude at unmeasured locations, this entry to the Spatial Interpolation Comparison of 1997 (SIC'97) used 2-dimensional, anisotropic, inverse-distance weighting interpolator (IDW), with cross-validation as a method of optimizing the interpolator's parameters. A jackknife resampling was then used to reduce bias of the predictions and estimate their uncertainty. The method is easy to programme, "data driven", and fully automated. It provides a realistic estimate of uncertainty for each predicted location, and could be readily extended to 3-dimentional cases. For SIC97 purposes, the IDW was set to be an exact interpolator (smoothing parameter was set to zero), with the search radius set at the maximum extend of data. Other parameters were optimized as follows: exponent = 4, anisotropy ratio = 4.5 and anisotropy angle = 40°. The results predicted by the IDW interpolator were later compared with the actual values measured at the same locations. The overall root-mean-squared-error (RMSE) between predicted and observed rainfall for all 367 unknown locations was 6.32 mm of rain. The method was successful in predicting 50% and 65% of the exact locations of the twenty highest and lowest measurements respectively. Of the measured values, 65% (238 out of 367 data points) fell within jackknife-predicted 95% confidence intervals, uniquely constructed for each predicted location.**

KEYWORDS: **cross validation, jackknife, uncertainty, IDW, anisotropic, automated, spatial interpolation, GIS.**

ACKNOWLEDGEMENTS The assistance of Ken Drouillard, Dr. Douglas Haffner and Dr. Alexander McCorquodale in reviewing this manuscript is gratefully acknowledged.

Contents

1. Introduction

In this article, an anisotropic inverse distance weighting interpolator (IDW) is used to make the required estimates of rainfall at 367 locations, based on a "training" set of rainfall measurements at 100 rain-gauges located throughout Switzerland. The method consists of two parts. First, the optimum set of parameters for the IDW method is selected via cross-validation and the estimates are made using this optimal interpolator. Second, the uncertainty of each of the estimates is calculated via the jackknife procedure.

The method is data-driven and fully automated (i.e., does not require preprocessing), which could be of value in an emergency situation requiring rapid yet justifiable results. Although the method attempts to optimize all the IDW parameters (i.e., weighting power, smoothing parameter, anisotropy ratio, anisotropy angle, and search radius) some parameters which are known or measured *a priori* (such as anisotropy angle determined by known wind direction), can either be fixed or limited to a user-defined range. This could result in improved performance and added realism of the interpolator. As applied in this paper, the technique fixes two and optimizes the other three out of five possible IDW parameters.

Compared with other methods, most notably kriging, the IDW method is simpler to programme and does not require pre-modeling or subjective assumptions in selecting a semi-variogram model (Henley, 1981). It provides a measure of uncertainty of the estimates that is directly related to the values being estimated, in contrast to kriging standard deviation which is based on the modeled semi-variogram (Adisoma and Hester, 1995). In addition, the IDW method is applicable to datasets of small size for which the modeled semi-variograms are very difficult to fit (Rasmunsen-Rhodes and Mayers, 1993), and it is flexible enough to model the variables with a trend or anisotropy present. The method is not limited to predicting rainfall measurements and it can be useful in problems as diverse as mapping of crop spraying, estimating grade and exploration feasibility of mining reserves, monitoring extend of contaminated groundwater plumes or quantitatively assessing the extent of contamination in aquatic sediments (Tomczak and McCorquodale, 1997).

2. Methods

The object of any two-dimensional interpolation is to estimate the value of a parameter (Z), at the unmeasured locations (Z_j) based on finite set of measurements of this parameter at other locations (Z_i). In SIC97 dataset, the parameter Z represents daily rainfall intensity. The IDW algorithm, as applied to each location being estimated, is based on Equation 1 (Keckler, 1995; Song and DePinto, 1995):

$$Z_j = \frac{\sum_{i=1}^{n} \frac{Z_i}{\left(h_{ij} + \delta\right)^{\beta}}}{\sum_{i=1}^{n} \frac{1}{\left(h_{ij} + \delta\right)^{\beta}}} \qquad (1)$$

Where Z_j is the interpolated value of a grid node, Z_i are the neighboring data points, h_{ij} are the distances between the grid node and data points, β is the weighting power, and δ is the smoothing parameter. In an isotropic case (i.e., when the weights are not a function of direction), Equation 1 can be used "as is" with the separation distance (h_{ij}) calculated by a simple Euclidean distance equation:

$$h_{ij} = \sqrt{(\Delta x)^2 + (\Delta y)^2} \tag{2}$$

Where: Δ x and Δy are the horizontal and vertical distances between the interpolated node "j" and the contributing data point "i".

In situations when inclusion of anisotropy is appropriate (such as in rainfall intensity, likely affected by wind direction and topography) the actual distance (h_{ij}) is replaced by the *effective distance ($h_{ij\text{-}eff}$)* which is calculated below (Keckler, personal communication, 1997). The equation is broken down for clarity.

$$h_{ij-eff} = \sqrt{A_{xx} \cdot (\Delta x)^2 + A_{xy} \cdot \Delta x \cdot \Delta y + A_{yy} \cdot (\Delta y)^2}$$

$$A_{xx} = T_{xx}^{\;2} + T_{yx}^{\;2} \qquad T_{xx} = \frac{\cos(\theta)}{\rho}$$

$$A_{xy} = 2 \cdot (T_{xx} \cdot T_{xy} + T_{yx} \cdot T_{yy}) \qquad T_{xy} = \frac{\sin(\theta)}{\rho}$$

$$A_{yy} = T_{yy}^{\;2} + T_{xy}^{\;2} \qquad T_{yx} = -\sin(\theta)$$

$$T_{yy} = \cos(\theta) \tag{3}$$

Where θ is the anisotropy angle (the direction of "preferred" anisotropic axis, counter-clock-wise from positive x-axis) and ρ is the anisotropy ratio (in isotropic case, ρ =1).

Conceptually, the effective distance can be thought of as shortening the distance between a data point and the interpolated node by the factor equal to the anisotropy ratio. The data point's relative influence on the interpolated node increases as the direction of line between the two points approaches the anisotropy angle. This concept is illustrated in Figure 1.

2.1 Cross-validation

IDW interpolator is driven by the set of parameters whose values are usually chosen at the operator's discretion. Parameters include:

- β - the weighting power (exponent)
- δ - the smoothing parameter
- ρ - the anisotropy ratio
- θ - the anisotropy angle

In addition, the search radius can be adjusted, determining the number of neighboring data points that are used when interpolating each node. Restricting the search radius can make the algorithm more efficient when the sample size is large and it can provide means to tackle a "trend" in the data (i.e., lack of stationarity).

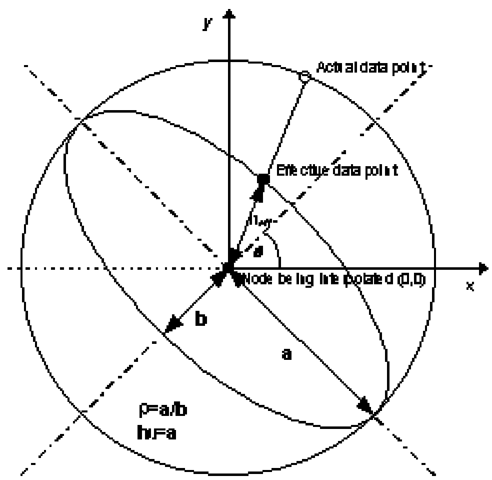

Figure 1. Illustration of the concept of anisotropy-corrected effective distance.

Searching algorithm can also be fine tuned by incorporating directional search (e.g., searching quadrants) or treatment of repeated measurements (e.g., averaging datapoints that are within some threshold distance). Although all of the above parameters can be adjusted, often some of them are known *a priori*, such as the anisotropy angle (but usually not ratio). Hence, based on the knowledge of the nature of data being sampled and processes involved (e.g., prevailing wind direction etc.) some IDW parameters can be fixed before the calculations start. Values of other parameters have to be selected and this choice greatly affects the results of the interpolation.

Although no measures are known that would or could be universally applied to choose the optimal set of parameters, cross-validation (a.k.a. "leaving-one-out" method) is often used to select an interpolator from finite number of candidates (Davis, 1987). The method is based on removing one data point at a time, performing the interpolation for the location of the removed point using the remaining samples (i.e., pretending that removed point does not exist), and calculating the difference (residual) between the actual value of the removed data point and the estimate for this point obtained from remaining samples. This scenario is repeated until every sample has been, in turn, removed. The overall performance of the interpolator is then evaluated as the root-mean of squared residuals (Davis, 1987; Song and DePinto, 1995).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Z_{i(int)} - Z_i)^2}{n}} \qquad (4)$$

where *RMSE* stands for root-mean-squared error, $Z_{i(int)}$ is the interpolated value of variable at point *i* estimated from remaining *n*-1 points, Z_i is the measured value of variable at the removed point *i*, and *n* is the number of data points.

Low root-mean-squared error (RMSE) indicates an interpolator that is likely to give reliable estimates for the areas where the rainfall intensity is not known. The cross-validation is performed with different set of parameters each time and the set with the lowest RMSE is taken as optimal. The step size and range of values for each parameter during the fitting procedure is user-specified. With IDW parameters selected via cross-validation, a jackknife resampling can be used to reduce bias of predicted values and to estimate their uncertainty.

2.2 Jackknife

In jackknife, all *n* samples (measured locations) are used to estimate parameter *Z* at interpolated node "*j*" (Z_{ALL}). The jackknife then proceeds by removing one observation at a time from the original dataset (one rain gauge from 100 known training locations), and repeatedly estimating value of parameter *Z* at node "*j*" from the remaining (*i-1*) data points. Let Z_{-i}be the corresponding estimate when i^{th} sample is omitted. A pseudo-value (Z^*_i) corresponding to each omitted point*,* is calculated as follows (Tukey, 1970):

$$Z_i^* = n \cdot Z_{ALL} - (n-1) \cdot Z_{-1}; \quad i = 1,2,\ldots,n \tag{5}$$

where *n* is the sample size, Z_{ALL} is the parameter estimate for node j using all *n* data points, Z_{-1} is the parameter estimate when i^{th} sample is removed, Z^*_i is a pseudo-value estimate for node "j" corresponding to i^{th} data point being removed.

The jackknifed estimator of parameter *Z* at location "*j*" is the mean of all pseudo-values for the node "*j*":

$$Z_J = \frac{\sum_{i=1}^{n} Z_i^*}{n} \tag{6}$$

where Z_J is the jackknife estimator of parameter *Z.*

The jackknife procedure is then repeated and the values of Z_J and σ_J are calculated for each estimated node location "*j*". The use of pseudo-values allows to assess the precision of the jackknife estimator through estimated standard error σ_J which is defined as (Adisoma and Hester, 1996; Efron and Gong, 1983):

$$\sigma_J = \sqrt{\frac{1}{n\cdot(n-1)} \cdot \sum_{i=1}^{n}\left(Z_i^* - Z_J\right)^2} = \sqrt{\frac{n-1}{n} \cdot \left(\sum_{i=1}^{n}\left(Z_i^*\right)^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Z_i^*\right)^2\right)} \tag{7}$$

If pseudo-values are treated as if there were *n* independent estimates, a confidence interval on the estimate can be constructed. Under the assumption that the statistic: (Z_J-

$Z)/\sigma_J$ follows Student's t distribution, a 100(1-α) % confidence interval on Z_J is given by (Adisoma and Hester, 1996; Efron and Gong, 1983):

$$Z_J \pm t_{\alpha/2,n-1} \cdot \sigma_J \tag{8}$$

where $t_{\alpha/2,n-1}$ is the is the value of CDF of Student-t distribution at 1-α confidence level, with n-1 degrees of freedom.

This approximation may not always be valid (Miller, 1964), but it should perform quite well providing the value of parameter Z does not depend essentially on only one or two sample points z_i (Tukey, 1970). Larger sample size will also improve the validity of such constructed confidence interval, courtesy of the Central Limit Theorem.

Figure 2 schematically shows the procedures involved in both cross-validation and jackknifing based on 5 hypothetical data-points for the location "j" being estimated.

3. Results

The interpolation method used here consisted of two steps: first the IDW parameters were optimized via cross-validation. To make the calculations more efficient and shorten the processing time, two parameters considered relatively less important were fixed *a priori*: the search radius was set to the maximum extent of the data and the smoothing parameter was set to zero (i.e., resulting in an exact interpolator since the uncertainty of individual rain-gauge measurements were assumed to be much lower than the uncertainty of the predictions). The remaining three parameters were simultaneously adjusted during the procedure (exponent: range 1 to 10, step 0.5; anisotropy ratio: range 1 to 10, step 0.5; anisotropy angle: range 0º to 170º, step 10°). The procedure produced the set of optimal (lowest cross-validated RMSE = 56.44) IDW parameters: Exponent = 4, Anisotropy Ratio = 4.5 and Anisotropy Angle = 40°. A jackknife was then used for each estimated point to reduce bias and to estimate the standard error and the resulting confidence intervals for each estimated point.

After the entries to the SIC'97 were submitted, the true measured rainfall values at all the 367 locations to be estimated were released to participants to allow for the evaluation of the performance the interpolators used. The performance of the IDW method along with the statistics of true measurements for all 367 estimated data points are listed in Table 1. The estimators of method's performance selected by SIC'97 organizers were: Root-Mean-Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Relative Error (MRE). Since the presented method has the ability of calculating the error and confidence interval for each estimated point, the summary of lower and upper 95% confidence levels for each calculated statistic, based on all predicted values, is also listed.

1

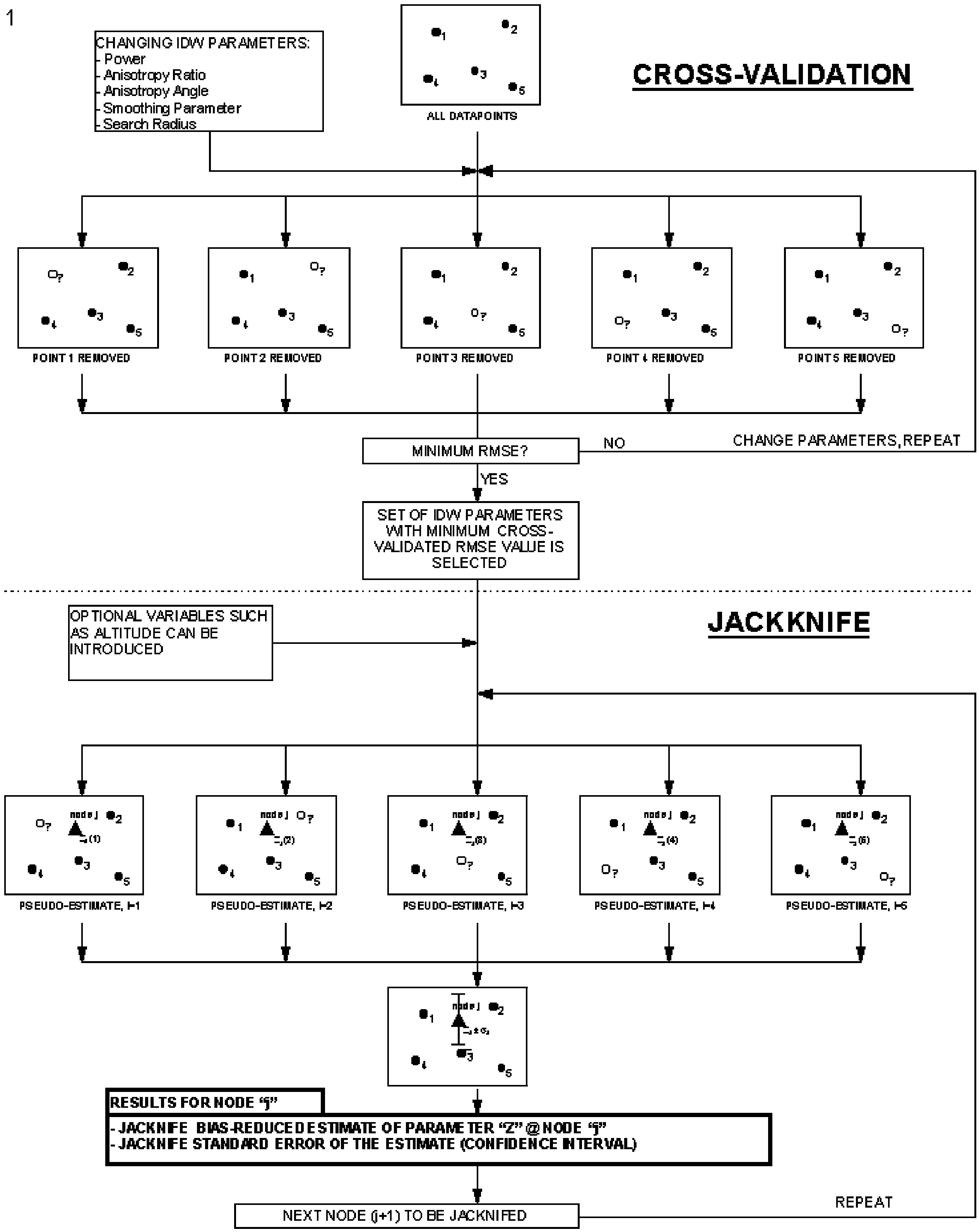

Figure 2 A schematic flow diagram of cross-validation and jackknifing for five hypothetical observed data points and the prediction made at node "*j*".

Table 1 Comparison of the estimated and measured values (n=367), 1/10 mm of rain.

Method	min	max	mean	median	s.d.	MAE	MRE	RMSE
Observed	0.0	517.0	185.8	162.0	111.2			
IDW (all points)	16.0	562.5	185.9	151.9	104.0	44.0	0.543	63.2
Jackknife "Corrected" IDW	0.0	787.6	185.3	145.2	127.0	58.5	0.565	83.9
Lower 95% Confidence Int.	0.0	344.1	117.2	99.5	93.4			
Upper 95% Confidence Int.	22.2	1275.8	264.4	209.2	173.3			

4. Discussion

As indicated in Table 1, the jackknife corrected estimates did not improve the performance of the interpolator for this particular application (higher RMSE than original prediction), and therefore the set of estimates using all training points (RMSE = 63.2) will be used in the remainder of the discussion. Jackknife is still used for estimates of uncertainty of IDW predictions.

Figure 3 shows the observed (true) rainfall measurements plotted against predicted ones for the same locations. The linear correlation coefficient of 0.83 confirms relatively good overall agreement (with no regard to spatial component in the data) between predicted and measured values.

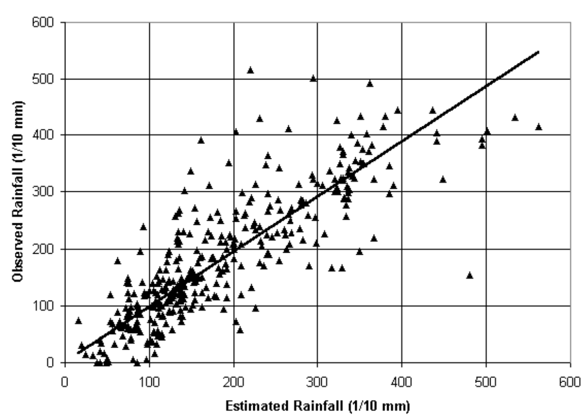

Figure 3. Observed vs. predicted rainfall for 367 "unknown" rain-gauges.

The overall predicted contour map is shown in Figure 4. The black circles show the relative magnitude of predicted rainfall (circle diameter) and location of all the 100 training sites from which the predictions were constructed. Contours exhibit a strong anisotropy in an approximate NE-SW direction consistent with the anisotropy ratio of 4.5 and anisotropy angle of 40° counter-clock-wise from the east (N60°E) used in the IDW interpolator.

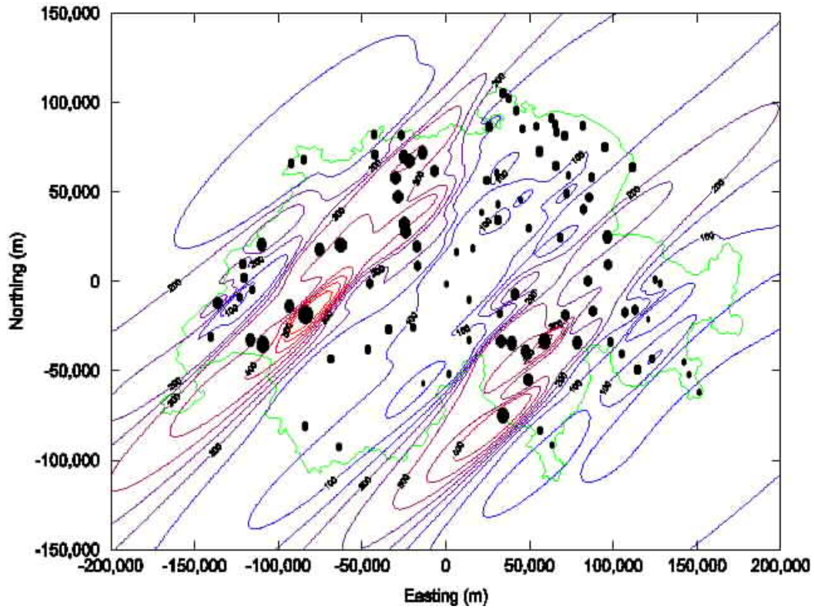

Figure 4 Predicted rainfall contours based on 100 given training points, superimposed on Switzerland's boarder, shown (size of dots is proportional to the magnitude of the recorded rainfall).

The bias of the predictions and independence of residuals from the magnitude of predicted values can be assessed from Figure 5. It shows the distribution of errors (value predicted minus value observed, or residual) as a function of the magnitude of observed values.

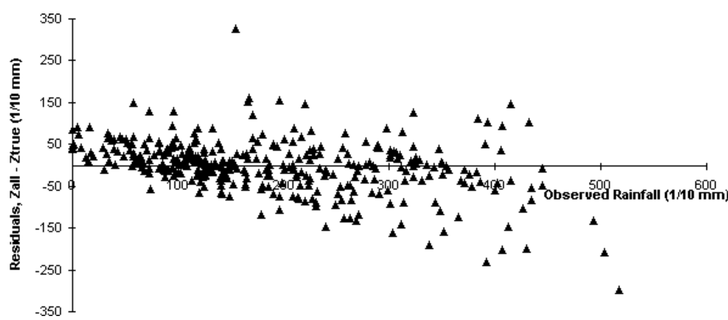

Figure 5. Distribution of residuals as a function of the magnitude of observed rainfall.

The residuals seems to have no overall bias (mean = 0.0, s.e.=3.3 , n = 363), with a weak tendency of being negatively correlated with the magnitude of observed rainfall (Pearson's R = -0.39). The *absolute value* of errors show weak positive correlation with the magnitude of the rainfall (R = 0.31). Although this linear correlation analysis does not capture the spatial nature of the dataset, it could indicate the possibility of a problem related to the *underprediction* of high values, which may be of an issue from the risk analysis perspective. The spatial distribution of errors (predicted versus observed rainfall values) is illustrated in Figure 6.

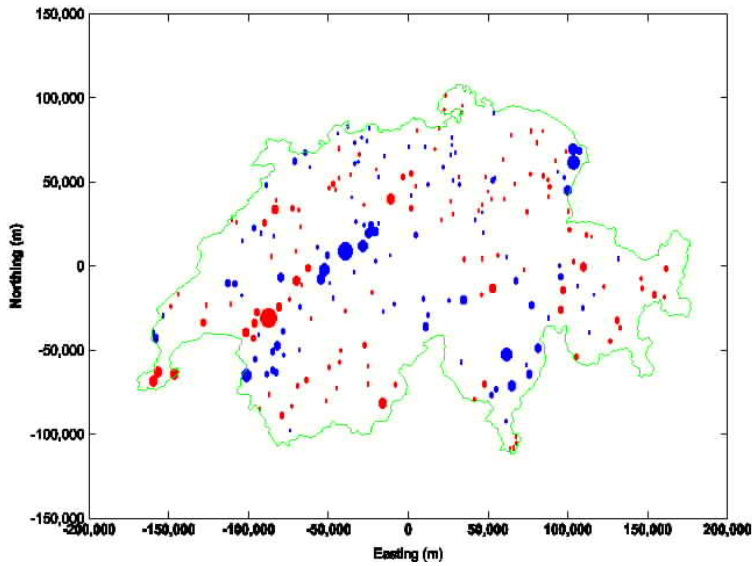

Figure 6 Error between observed and predicted values (predicted minus observed) with the size of the dots proportional to the absolute value of the error. Blue circles signify negative errors (under-prediction) and red circles indicate positive errors (over-prediction).

Figure 7 shows the distribution of predicted uncertainties of estimated values as calculated from Equation 7 (i.e., standard errors calculated via jackknife, for each estimated location).

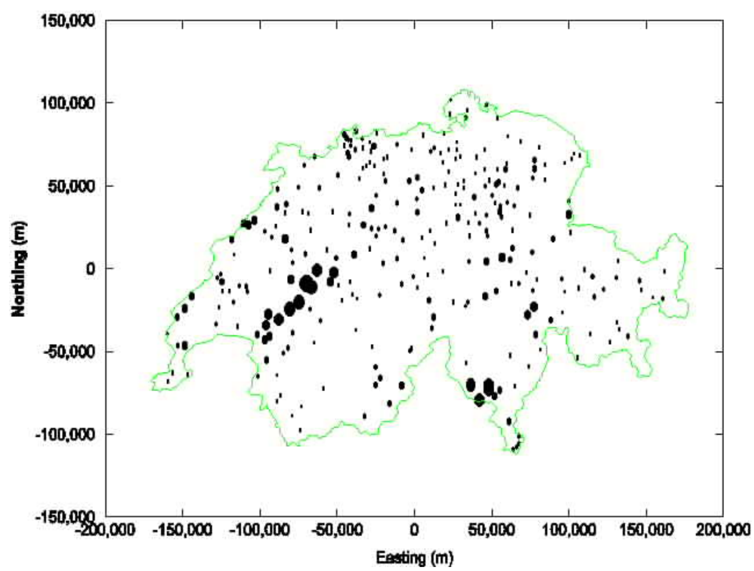

Figure 7 The value of the calculated jackknife standard error for each predicted point. Size of symbols is proportional to the magnitude of the error and uses the same scale as Figure 6.

When Figure 6 and 7 are compared, the overall pattern and magnitude of predicted uncertainties and observed errors is similar, with the highest uncertainty (for both actual errors and predicted uncertainties) concentrating in the west and south part of the country. Of the measured values, 65% (238 out of 367 data points) fell within 95% confidence interval as predicted for each estimated point with Equation 8 (i.e., based on jackknifed standard error for each predicted point).

The ability to predict extreme values is an important benchmark in evaluating the performance of an interpolator. Some of the highest measured rainfall values were included in the set of 100 training data points and, therefore, were not estimated via interpolation. To facilitate a fair discussion on the method's performance, only the extreme values in the predicted dataset (367 data points) are included in the following section. Figure 8 shows the distribution of ten highest and lowest rainfall values as requested by SIC97 organizers (both predicted and observed as indicated in the legend).

Of the extreme values, 20% of the highest and 30% of lowest values were predicted accurately with the overall pattern and approximate location for predicted and observed extremes being similar. It was felt that the comparison based on the arbitrary 10 extreme values (only 2.7% of all the values estimated) may not be representative for measuring the method performance. If a similar comparison of the predicted versus observed extremes is made based on 20 highest and 20 lowest points (Figure 9), the efficiency of correctly predicting highest and lowest locations increases to 50% and 65% respectively.

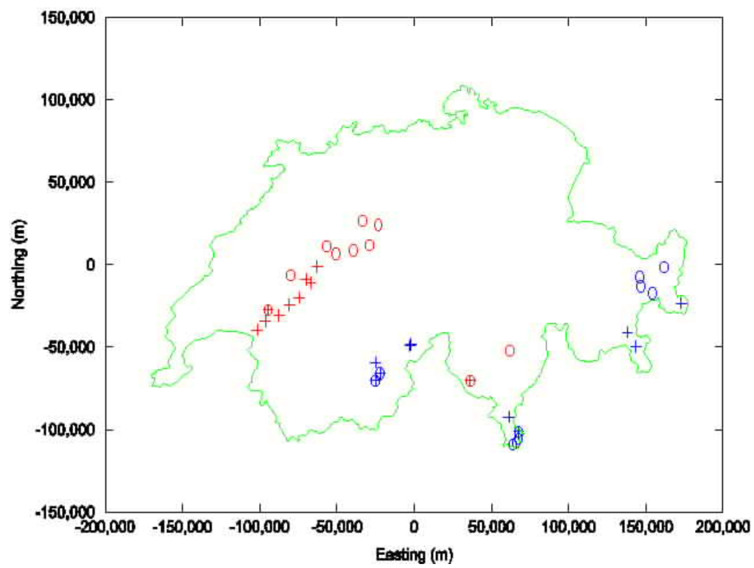

Figure 8 Ten highest and ten lowest rainfall values. Red symbols signify high values, blue symbols show low values. Circles represent observed values, crosses represent predicted values.

Taking into account the fact that the rainfall data set used in the SIC'97 comparison was related to the Chernobyl accident (which was not known by participants before the entries were submitted), the interpolators allowing for rapid estimates without the need of preprocessing are to be preferred. The automated IDW interpolator appears to satisfy this mandate.

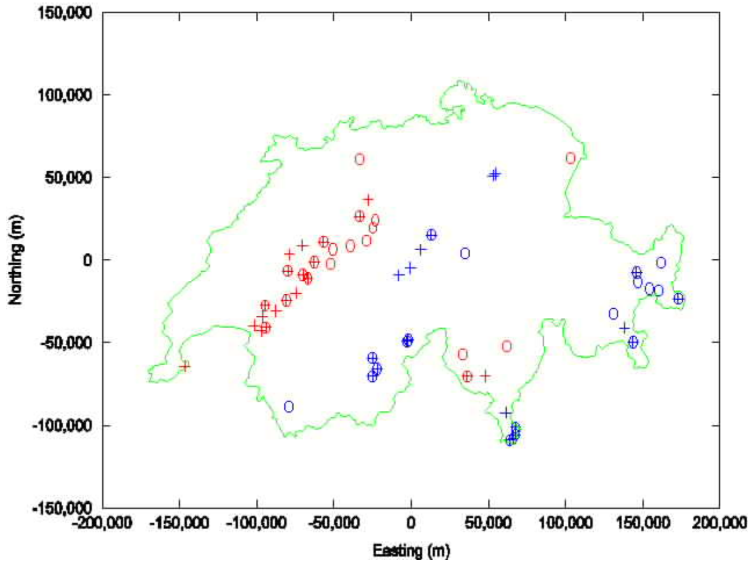

Figure 9 Twenty highest and twenty lowest rainfall values. Red symbols signify high values, blue symbols show low values. Circles represent observed values, crosses represent predicted values.

The anisotropic IDW with cross validation and jackknife is certainly not a "silver bullet" for contouring all spatially distributed variables. As with any spatial interpolation method, one could concoct situations for which models other than IDW were more applicable. It appears, however, that considering its ease of programming, automation, flexibility, objectivity, ability to measure the uncertainty of the predictions, and a good performance of the IDW model with this and other data sets, the method can be considered a sound, robust, general purpose 2D interpolator. The method could also be readily extended to 3-dimentional cases.

The modified IDW method, as described in this paper, allows for fast estimates even with moderate computing power (Pentium133 PC), as well as for assessing the uncertainly of these estimates. The algorithm can be programmed as a stand-alone application or as a part of a GIS Decision Support System. For the purpose of SIC'97, all the procedures were programmed in GS-Scripter (a modified version of BASIC) and used IDW interpolation subroutine included in Surfer3D by Golden Software Inc. (Keckler, 1995), via OLE 2.0 automation.

References

Adisoma, G. S. and M. G. Hester (1996) Grade estimation and its precision in mineral resources: the Jackknife approach. *Mining Engineering*. 48(2); 84-88.

Davis, B. M. (1987) Uses and abuses of Cross-validation in geostatistics. *Mathematical Geology,* 19(3); 241-248.

Diaconis, P. and B. Efron (1983) Computer-intensive methods in statistics. *Scientific American*, 5.

Efron, B. and G. Gong (1983) A leisurely look at the Bootstrap, the Jackknife, and Cross-Validation. *Am. Stat.* 37(1); 36-48.

Efron, B. and R. J. Tibshirani (1993) *An Introduction to the Bootstrap*. New York, NY.: Chapman and Hall, Inc.

Henley, S. (1981) *Nonparametric Geostatistics*. London: Applied Science Publishers, Ltd.
Keckler, D. (1995) *The Surfer Manual*. Golden, CO.: Golden Software, Inc.
Miller, R.G. (1964) A trustworthy jackknife. *Annals of Mathematical Statistics*. 35; 1594-1605.
Rasmunsen-Rhodes, H. and D. E. Mayers (1993) GRASS Used in the geostatistical analysis of lakewater data from the Eastern Lake survey, Phase I. In M. F. Goodchild, B. O. Parks and L .T. Steyaert (eds.): *Environmental Modeling with GIS*, M.F. Oxford: Oxford University Press, Inc.
Song, J. and J. V. DePinto (1995) A GIS-based Data Query System. Presented at the International Association for Great Lakes Research (*IAGLR)* Conference, Windsor, Ontario.
Tomczak, M. and J. A. McCorquodale (1997) GIS Supported Mapping and Modeling of the St. Clair River Contaminated Sediments. Presented at the International Association for Great Lakes Research (*IAGLR*) Conference, Buffalo, NY.
Tukey, J. W. (1970) Some further inputs, In: *Geostatistics*: *A Colloquium*, D.F. Merriam (ed.), New York, NY.: Plenum.

JGIDA vol. 2, no. 2
JGIDA Home